全生命周期视域下人文社科研究数据管理平台的设计与实现*

■ 姚占雷[1] 谷俊[2] 许鑫[1,3]
[1] 华东师范大学经济与管理学部 上海 200062 [2] 上海师范大学人文学院 上海 200233
[3] 华东师范大学调查与数据中心 上海 200241

摘 要：[目的/意义]近年来，我国虽已出台政策对科学数据的管理、共享与利用给出了明确的引导与规范，但现有主流的数据管理平台架构侧重在对数据的科学管理，数据的共享利用效率不高，本研究在系统拓展现有平台的数据管理功能基础上，聚焦解决科学数据的共享利用难题。[方法/过程]在广泛调研与文献梳理的基础上，首先厘清人文社科研究数据管理平台建设的必要性和困境，其次系统设计与阐释了全生命周期视域下人文社科研究数据管理平台建设的核心功能与特点，进而结合平台实例，详细描绘了核心功能的关键技术实现。[结果/结论]以开放互联为基础、以开发利用为核心、以自助分析为特色，最终建立起面向全生命周期的研究数据管理平台基础框架，并据此设计实现了一个全生命周期的人文社科研究数据管理平台，可为相关实践与研究提供特色案例与参考。

关键词：数据管理 人文社科 开发利用 平台建设
分类号：G252.5
DOI：10.13266/j.issn.0252-3116.2021.07.003

科学数据是国家科技创新和经济社会发展的重要基础性战略资源，在当今大数据时代，科技创新活动越来越依赖于对科学数据的分析挖掘和综合利用。为此，我国出台了《科学数据管理办法》(国办发〔2018〕17号)，强调要"加强和规范科学数据管理，要适应大数据发展趋势，积极推进科学数据资源开发利用和开放共享"。不过研究数据共享与重用的价值虽在学界已成共识，但在实践上不尽如人意[1]，同时作为一类基础性资源，我国人文社科数据资源建设相对滞后且多由国家重大科研项目驱动，并因"数据服务平台功能单一、检索效率低下、不支持机器读取和原始下载、系统平台整体上可用性较差"等现实困境，而难以满足项目之外的用户需要[2]。目前，针对人文社科研究数据的科学管理与开发利用活动，也尤其欠缺。

随着学术研究的深入和跨学科科研方向的拓展，不同学科的科研人员将自身的科研经验、拥有的知识和各种研究数据进行共享，可为其他科研人员的研究提供更多的思路和灵感，进而提高科研人员借助跨学科和跨领域知识进行科学研究的能力[3]。为促进科研人员数据重用，相关研究人员已从数据管理者、数据重用者两个视角展开了大量研究与实践[1]，本文基于数据管理者视角，遵循"建好、管好、用好"建设思路，进一步完善与改进现有人文社科研究数据管理基础设施，旨在解决研究数据的共享利用难题。

1 相关研究与实践

1.1 研究数据管理

研究数据泛指科研活动中原始的、基础的数据，能够帮助提高科学的可再利用性和可信度，对它的管理应该是针对研究数据的整个生命周期的管理[4]，继而高效开展各类数据管理活动。当前，国内外研究人员围绕科研人员需求调查[5-7]、数据管理生命周期[8-10]、数据管理服务[11-13]、数据管理政策[14-16]、数据管理教育[17-19]等主题，做出了大量研究与探索，以美国、英

* 本文系华东师范大学"幸福之花"先导研究基金项目"大数据视阈下基于学术共同体的人文社科学术评价与促进研究"(项目编号：2019ECNU-XFZH016)研究成果之一。
作者简介：姚占雷(ORCID：0000-0002-7143-4725)，工程师，硕士；谷俊(ORCID：0000-0002-9385-0527)，副教授，博士；许鑫(ORCID：0000-0001-7020-3135)，教授、博士生导师，副主任，博士，通讯作者，E-mail：xxu@infor.ecnu.edu.cn。
收稿日期：2020-10-26 修回日期：2021-02-07 本文起止页码：25-37 本文责任编辑：杜杏叶

国、澳大利亚为代表的研究数据管理研究与实践活动起步较早，已形成了与各自国家科研文化背景相适应的不同发展路径和解决方案[20]，且高校图书馆的作用和地位愈加凸显[3]。而国内研究多是介绍国外研究数据管理研究与实验经验[21]，并在此基础上开展系列实践活动，尤其是2011年后图书情报学界和图书馆行业开始主动介入、跟踪和开展研究数据管理研究与实践活动，出现了以中国科学院"科学数据管理与共享云服务平台"、武汉大学图书馆CALIS三期的"高校科学数据管理机制及管理平台研究"等为代表的典型实践案例。不过，我国的研究数据管理研究与实践活动还处于探索发展阶段，良性的开放共享文化和机制尚未形成，系统理论研究和过程评估方法的全面、整体、启发式梳理创新不足[18]，相关研究亟需加强。

1.2 研究数据生命周期管理

研究数据不同于信息资源"价值老化"的生命周期衰变规律，其生命周期与科学研究活动联系紧密，且受到研究方法、工具、手段等的影响，是研究如何在数据生命周期各个阶段采用适当的操作与策略对数据进行管理，其管理的对象除了数据本身，也包括数据的生产、服务、使用对象和内外部环境、技术等[22]，它具有数据本身的生命周期管理和反映科研活动生命周期的两重性[9]。

然而，尽管已有学者基于科研活动生命周期探究数据资源的整合路径，如井润田等[23]聚焦科研活动中的团队，剖析了科研团队不同生命周期特点，在此基础上，贾玉文等[24]建立了嵌入科研生命周期的资源整合模型，但目前主流常见的数据管理生命周期模型[25]主要还是面向数据本身（见表1）展开，相关研究实践活动也多在此基础上开展，同时为有效测度、评估和持续改善研究数据管理的实践与服务，构建有研究数据管理能力成熟度模型[26]，进行分级测度。

表1 几种典型的研究数据管理生命周期模型

数据生命周期模型	要点摘录	提出机构/个人(年)
DCC	6个阶段：概念化，创建和接收数据，评测和选择数据，长期保存和存储，访问、使用和重用，转换	英国数据管理中心(2004)
DDI	8个阶段：概念研究、数据采集、数据处理、数据存档、数据发布、数据发现、数据分析和数据重用	英国数据档案项目联盟(2014)
DataONE	8个动词：计划、收集、保证、描述、保存、发现、整合、分析	美国新墨西哥大学图书馆等(2009)
UKDA	6个阶段：数据创建、数据加工、数据分析、数据保存、数据访问、数据再利用	英国埃塞克斯大学(2007)
ANDS	8个动词：创建、存储、描述、识别、注册、发现、获取、开发	澳大利亚国家数据服务(2008)
I2S2	2个阶段：基础阶段（提出计划，同行评议，进行实验，数据处理、分析和解释，最终报告研究成果）和理想化阶段（评估和质量控制，元数据和上下文信息的文件，存储、归档、保存和管理、知识产权、禁止和访问控制）	英国结构化科学整合基础设施项目(2009)
OAIS	6个功能实体：数据收集、归档存储、数据管理、管理、保存规划和数据访问	N. Beagrie等(2001)
Research360	6个阶段：计划和设计、收集和获取、解读和分析、管理和保存、发布和出版、挖掘和再利用	英国巴斯大学(2013)

表1所示的研究数据管理生命周期模型，主要描述了如何做好数据的管理与控制，其中虽然提及到数据的重用与开发利用，但并未细化、展开，且多是面向机构（管理者）。而对研究数据的使用包括了数据的验证、聚合、挖掘、再利用四方面[27]，它能够进一步促进学术新发现、形成学术新生态等，因此对数据管理生命周期模型的优化，有其必要性。

1.3 研究数据管理平台与工具

针对研究数据管理的服务供给，需要依赖平台和工具，并据此处理研究数据管理生命周期各个环节中的数据管理问题（见图1），现阶段围绕着研究数据的管理，数据管理平台和工具出现"百花齐放，百家争鸣"的状态，且朝着开放、融合、标准化的方向发展[28]。然而，图1所示目前主流的数据管理平台工具多是侧重在数据的创建、处理、保存和访问环节，针对分析和重用环节的并不多见。

（1）数据管理计划工具。主要是对数据管理进行概要性描述的正式文件，它包括了项目进行过程中及项目完成后等各个阶段[29]。目前，影响最大、使用最广泛的数据管理计划工具（Data Management Plan，DMP）主要有三个，依次为DMPonline（https://dmponline.dcc.ac.uk）、DMPTool（https://dmptool.org）和DMP Roadmap（https://github.com/DMPRoadmap），且均为开源软件。

（2）实验室电子笔记。主要是将实验数据以电子的形式记录存储，并提供协作、模板、数据收集与分析等功能，以提升研究流程优化和过程记录。美国明尼苏达大学图书馆曾于2017年开展了一项针对美国顶尖研究大学实验室电子笔记应用情况的专项调查，结果显示绝大多数实验室电子笔记的价格昂贵，且已有

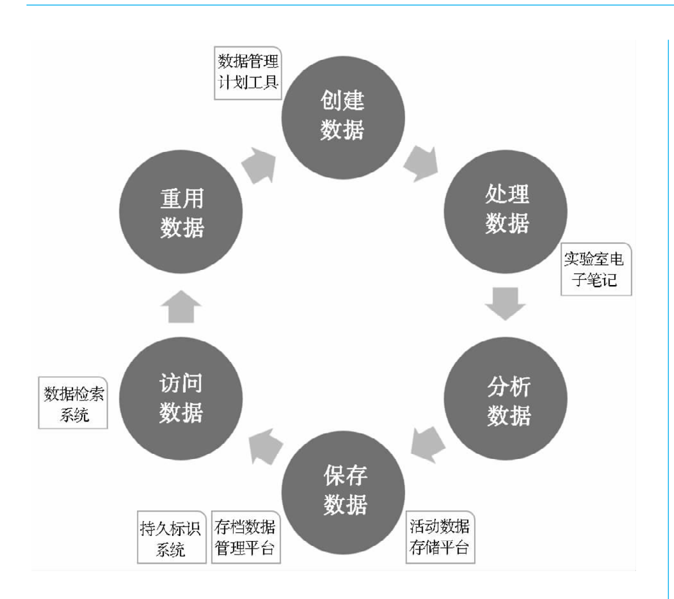


图 1　数据管理平台工具及其在典型的数据管理生命周期中的分布示意

图书馆开始提供实验室电子笔记服务[30]。目前，可供使用的实验室电子笔记较多，但因研究的差异软件，没有一个实验室电子笔记能够满足所有研究者需求，一些常见的实验室电子笔记软件有 LabArchives（可试用，分为专业版和教学版）、RSpace（分为社区版和企业版，其中社区版可免费试用）、sciNote（分为免费版、高级专业版和高级企业版，是开源软件）等。

（3）活动数据存储平台。在科学研究过程中，研究者会不断地产生数据，这些数据通常称为"活动数据"，数据的安全防范（涉及硬件损害、病毒入侵、误删除等）至关重要。随着云计算技术成熟与普及，针对此类数据的存储，除了传统的多重备份、异地备份外，也多了选择、渐趋"上云"，如：选择通用的公有云存储（Google Drive、百度网盘等）、购买商业服务搭建校园云存储以及利用开源软件自建云存储（针对高安全等级的数据要求）。

（4）存档数据管理平台。即传统意义上的研究数据管理平台，用于管理那些高稳定性、重要的且需长期保存的研究数据，如：自建的 ICPSR，开源的 Dataverse、Dspace，商业的 Figshare 等平台。目前，已建成在用的存档数据管理平台非常多，涉及众多学科领域，同时专门出版数据的平台业已出现，如：自然出版集团推出的 Scientific Data（2014）、《全球变化数据学报（中英文）》编辑部推出的全球变化科学研究数据出版系统（2014）、《图书馆杂志》编辑部推出的数据管理平台（2017），等等。re3data. org 网站显示，截至 2021 年 1 月 13 日，已经注册的平台已达 3 581 个。其中，美国居首位有 1 100 个，剔除国际协会/组织（249 个）外，排名前十的国家拥有 2 762 个，中国仅有 47 个，与美国等排名靠前国家相比，差距显著。

（5）持久标识系统。是指为数据分配全球唯一、持久的标识符，以便于数据资源的引用、识别、定位和长期保存。目前，在数据管理平台被广为使用的持久标识符方案主要有三种，即：Handle（http://www.handle.net）、DOI（http://www.doi.org）、ARK（https://n2t.net/e/ark_ids.html）。

（6）数据检索系统。是用于支撑研究者找到研究所需的数据资源，分为数据集检索系统（直接对数据集本身的元数据检索）和数据仓储检索系统（侧重对数据仓储的元数据检索）。目前，主流常见的数据集检索系统有 Data Citation Index（商用）、DataCite Search、Google Dataset Search，数据仓储检索系统有 re3data、FAIRsharing。

1.4　人文社科研究数据管理平台现状

有别于自然科学，人文社科更关注人文社会现象及其规律性的系统认识、具有社会意识性质，它虽数据规模小，但蕴含的语义内容丰富多样，且具有高度可复用特性，即人文社科数据的使用周期较长，同一研究方向的社科数据可以被多个研发团队复用，数据可以产生持续的价值。然而相较于自然科学领域，我国人文社会研究数据无论是原始科研数据还是衍生数据，大多止于对应的学术成果发表后，共享与深度开发利用的环境尚未真正形成，学者数据共享与利用积极性仍有待提高，且人文社科数据资源建设相对滞后且多由国家重大科研项目驱动，并因面临着"数据服务平台功能单一、检索效率低下、不支持机器读取和原始下载、系统平台整体上可用性较差"等现实困境，而难以满足项目之外的用户需要[2]。

随着当前科学研究范式朝着数据驱动方向转型，以人文社科为代表的研究数据管理平台建设受到广泛关注，并涌现出了诸多案例（见表 2）。尽管我国的相关平台建设虽已取得较大成就，但人文社科领域正处于起步探索阶段、数据集欠规范且量少，且相较于国外，普遍存在着软件开发缺少开源理念、平台服务功能不全面、部分平台缺乏合作建设理念等问题[31]。而近年来出现的"全国高校数据驱动创新研究大赛""'慧源共享'高校开放数据创新研究大赛""'大师杯'数据联赛"等各类数据竞赛活动，虽为研究数据资源的二次开发利用提供了新的路径与尝试，但相关平台对数据分析的支撑能力（算力、工具包等），仍处于起步阶段。

表 2 几种典型的人文社科研究数据管理平台

平台名称	数据集个数	主要功能及官方网址	平台软件	国别
英国数据档案馆 UKDA	8 100	数据集的创建、提交、查找、下载,咨询服务,讨论社区 https://www.data-archive.ac.uk	Nesstar	英国
美国密歇根大学校际政治和社会科学研究联盟 ICPSR	15 600 +	数据集的创建、提交、查找、下载,数据分析,咨询服务,新闻事件发布,讨论社区 https://www.icpsr.umich.edu/icpsrweb	自主研发	美国
美国哈佛大学 – 麻省理工学院数据中心 HMDC	106 870 +	数据集的创建、提交、查找、下载,在线数据统计分析,支持研究计算,桌面服务和托管服务 https://dataverse.harvard.edu	Dataverse	美国
北大开放研究数据平台	300 +	数据集的创建、提交、查找、下载,在线数据统计分析 https://opendata.pku.edu.cn	Dataverse	中国
复旦社会科学共享数据平台	770 +	数据集的创建、提交、查找、下载,数据分类统计 http://dvn.fudan.edu.cn	Dataverse	中国
人大中国学术调查数据资料库	800	数据集的创建、提交、查找、下载,数据分析报告分享 http://www.cnsda.org	自主研发	中国

注:数据集个数的统计日期为 2021 年 1 月 13 日,未区分学科

需要说明的是,现有的研究数据管理平台多是将数据视作一类信息资源开展建设,虽然在顶层设计时考虑了数据的分析、挖掘与利用,但具体应用时重心不在于此。当下,数据资源的价值日益重要、二次开发利用需求日趋强烈,这需要对现有的研究数据管理平台进行升级改造。

2 人文社科研究数据管理平台设计

当前我国有关人文社科研究数据管理平台的建设,多是由高校主导建设,在数据内容上各具特色但欠规范、平台功能上二次开发利用不足。本文则聚焦数据资源的二次开发利用,遵循"建好、管好、用好"建设思路,在兼顾主流数据管理平台基础功能的基础上,围绕数据资源的多途径采集与规范管理、自助分析与开发利用,形成特色建设方案,如图 2 所示:

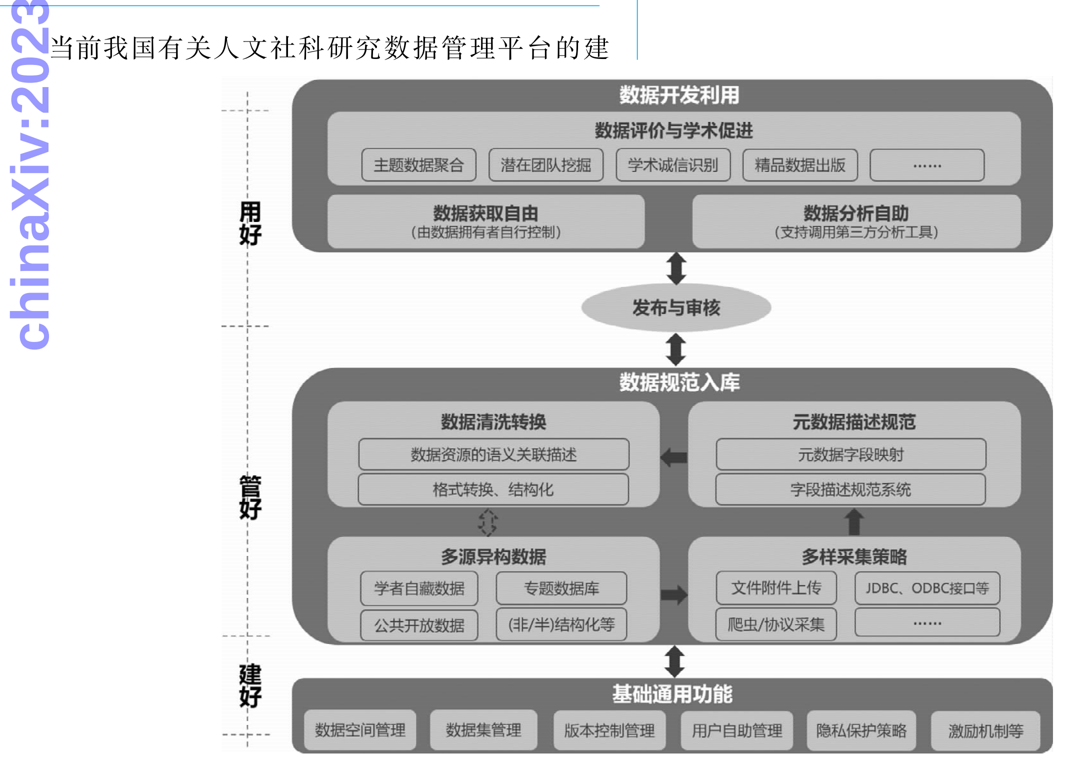


图 2 面向全生命周期的人文社科研究数据管理平台基础框架

在以上框架指导下,人文社科研究数据管理平台建设应注重在以下三个方向发力:

(1)开放互联的数据共享机制。重视数据管理平台的数据共享能力与安全保障,并在遵守相关法律法规、满足数据隐私包含相关条款下,实现数据资源的科学顺畅流动和有效利用。因此,平台建设过程中,应兼

顾数据接入（采集、交换等）、规范（元数据管理、备份等）、利用（聚合、分析、下载等）等在不同场景下的适用性，以保障数据资源在平台内外部的科学流动，同时考虑引入区块链等技术完成数据的确权与版权保护，引入沙箱等技术保障数据“不落地”的开发利用。

（2）自助分析的软件接入规范。重视数据管理平台的分析能力，为平台赋能，提升平台可用性，促使平台由“重藏”向“藏用”转变。因此，在常见的资源统计、预览等功能基础上，平台还需为研究人员解决平台繁杂数据资源的拼接利用问题，进而促进研究人员共享研究数据。为保障平台的分析弹性和扩展性，应重点解决第三方分析软件的自由接入问题，而不是自建分析环境，着重关注平台的数据分析接口规范开发，形成软件工具的开放互联机制，满足研究人员多样化的分析需要。

（3）数据资源的开发利用模式。重视平台数据资源的二次开发利用活动，在不违反相关法律法规的前提下，开展系列增殖服务（如研究数据的追踪与验证、聚合、挖掘、再利用等），发挥研究数据资源的重用价值。结合数据利用性质的迥异，平台建设应关注对数据集评价与学术促进、数据出版等活动的支撑作用，应具备面向主题、多源潜在关联数据的聚合能力，潜在科研团队的挖掘能力，研究数据追踪与学术诚信的识别能力，优质专题数据集（数据质量、利用率等）的洞察能力，等等。

2.1 数据采集与元数据管理

解决数据采集的多样化和元数据描述规范问题，平台应能够满足研究人员自藏数据、专题数据库、公共开放数据等不同来源数据的采集需求，以适应研究数据广泛散在分布的特点，并支持对数据的统一描述。

（1）数据多样化采集策略。研究数据管理平台中数据来源的性质，分为研究人员提交（专题数据）、平台主动采集（公共开放数据/委托加工数据）两种方式。但无论何种方式，得到的数据均需要高效稳定、精准无误地采集汇入平台，而数据本身除了传统经典的文件附件上传外，还会大量以数据库、网页等形式存在，且此类数据通常是海量的、可变的。为此，平台应革新采集策略，设计 JDBC/ODBC 接口、爬虫/协议采集、API 调用等，以适应这一现状，同时兼顾数据采集的数量与频次、数据呈现方式（统计图表、数据列表等），且具有良好的用户体验。

（2）元数据描述规范。针对研究数据特点，建立完整的元数据描述规范，实现对研究数据的标准化描述（如基本属性、特色属性、价值属性等三大属性，见图3），包括了字段描述规范系统和元数据字段映射。字段描述规范系统规定了平台数据发布的字段描述规则，对数据发布者起到指导作用，以提高平台其他用户对数据的理解度，提升数据的共享能力。

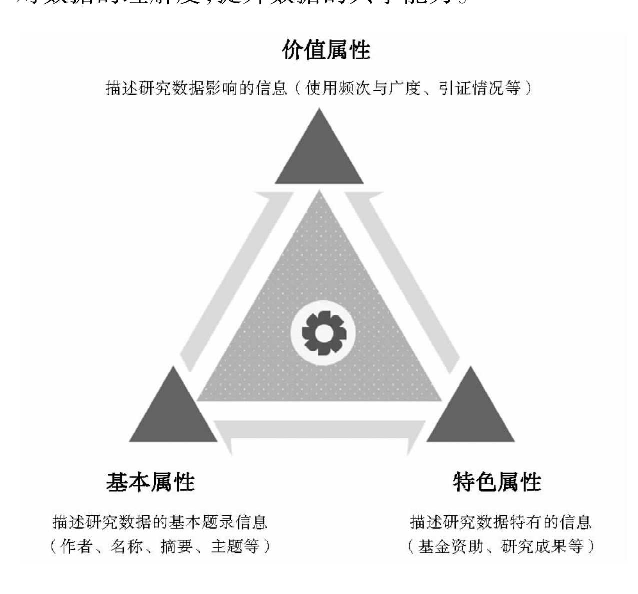


图3　研究数据的标准化描述示意

2.2 数据共享及其版权保护

解决研究数据的科学管理与共享问题，主要围绕研究数据的数据题录、数据文件两个方面来展开，旨在促进数据在平台内外部的有机流动。

（1）数据题录的管理与共享。数据题录摘自研究数据中的元数据，为支撑研究数据的最大化共享传播，平台应支持通过 RSS 订阅、一键分享、API 等方式将其分享，由此以主动方式对外推送、提升数据资源曝光度，同时遵循主流常见的数据互操作协议（如 OAI-PMH 协议、SRW/U 协议、SDARTS 协议等），满足跨平台的数据资源整合与共享。

（2）数据文件的管理与共享。关注研究数据本身，针对多源异构的各类型数据（.txt、.xlsx、.csv、.sql 等），不应仅停留在数据被提交到平台上，需注重数据的平台化应用，设计并实现一套融合多类型文件的存储机制（见图4），实现数据文件在内容级上的深度融合，由此为数据资源的二次开发利用、面向主题的多源数据聚合关联等，奠定基础。同时，为充分保障数据隐私等，针对数据文件的开发利用与共享活动，须征得数据所有者的同意，如：在数据正式发布前，提示数据所有者设置相应权限，而当数据文件的权限被设定为“受限”，在后续的开发利用与共享活动中如涉及该数据文件时，平台须通过邮件、短信等方式通知到数据所有者，以获取相关授权。

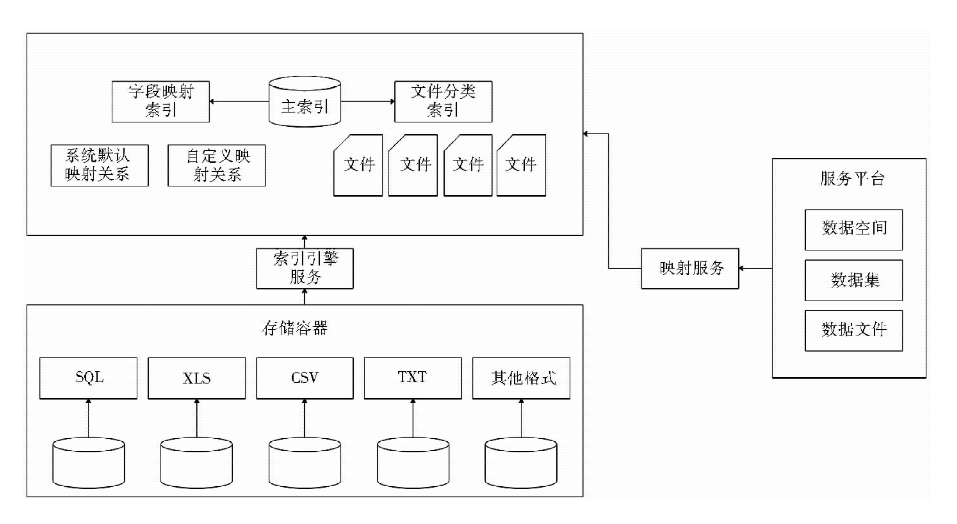


图 4　一种面向研究数据管理的融合多类型文件的存储机制

区块链的去中心化、开放性、自治性、信息不可篡改性、匿名性等特点，在人文社科数据共享过程中的自治、追踪、溯源等方面有着天然的优势[32]。考虑到公有链是面向所有公众开放、其数据的管理不受任何个人和组织的控制，且人文社科数据具有可持续使用、更新速度较慢等特点。因此，在人文社科研究数据管理平台建设中，建议采用私有链或联盟链进行数据确权与版权保护，且相较于公有链，节点间的交易成本低，如图 5 所示：

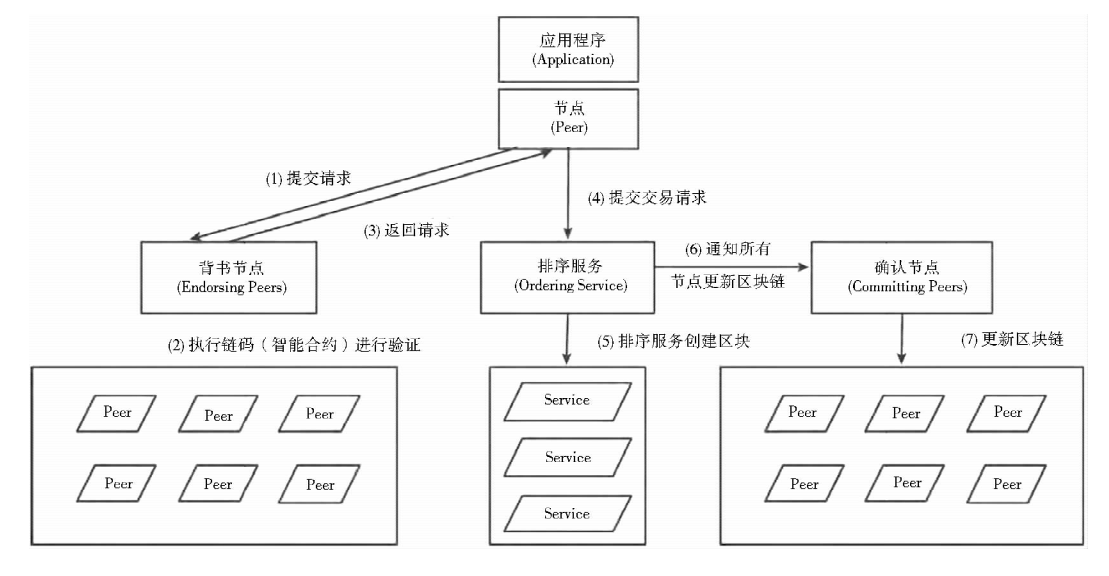


图 5　联盟链典型代表 Hyperledger Fabric 框架的节点间交易步骤示意

2.3　研究数据在线自助分析

面向研究人员，解决平台中研究数据的使用问题，助力研究人员深入内部观测数据详情、挖掘分析等活动，主要体现在兼容主流的分析软件工具上。

（1）通用的数据探索。应能满足研究人员针对感兴趣的数据文件进行浅层的数据概览需要，以更为直观把握数据的形态、质量、内容等，包括但不限于对数据文件的字段描述、数据实例、统计报表等信息。其中，为进一步辅助研究与观察，统计报表设计建议采用图 6 框架展开：

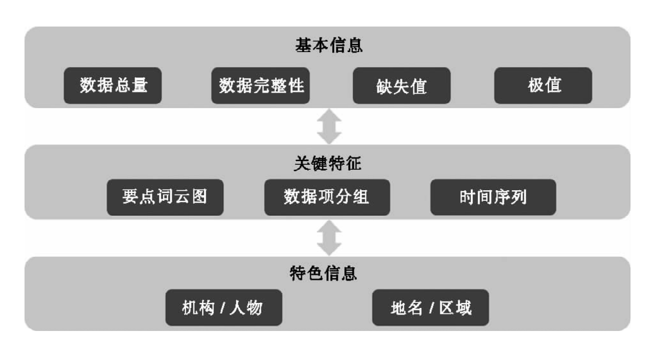


图 6　统计报表设计要点

（2）专业的挖掘分析。平台分析工具的丰富性，

既能支持研究人员围绕感兴趣的数据文件开展专业化的挖掘分析活动,亦可激发研究人员共享研究数据、增强平台活跃度。图 6 在数据层面解决了平台内多源数据表达问题,此处需要解决的是平台内数据迁移和工具接入问题。其中,平台内数据迁移问题是指在挖掘分析活动时平台数据是否要流向工具,关心的是执行效率,即数据向工具流动本身是需要时耗的,这在数据海量情况下极为重要;工具接入方式则分为硬接入和软接入两种,硬接入是与平台深度集成、融为一体,这不涉及数据迁移,软接入则是通过 API 调用等完成,多涉及数据迁移但可支持第三方工具自由接入,如图 7 所示:

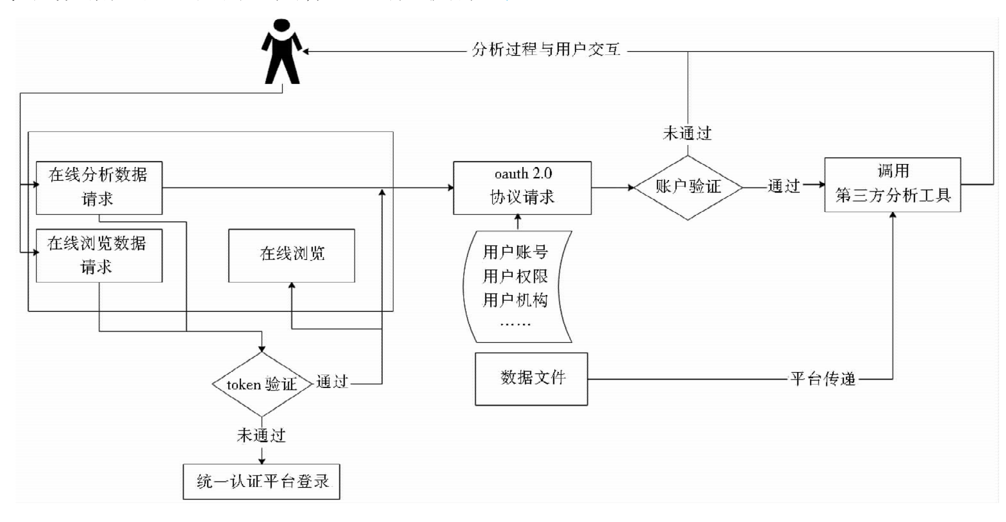


图 7　第三方工具接入平台的机制

2.4　数据资源二次开发利用

面向管理,助推平台中研究数据的二次开发利用活动,有针对性地开展数据增值服务,平台需要为这些服务活动的开展提供系列技术支撑。

(1)面向主题的数据聚合。主要依托元数据描述、数据文件中数据项的特征等,对来源于不同项目或不同研究人员的数据进行标签抽取与关联、主题相似测量、多维聚合,意在克服来源于单一数据可能存在的偏差。平台虽面向单一数据文件在数据层面做了统一规范,但面对研究数据本身的繁杂多样特点,还需提供支持研究人员进行多源数据拼接、关联的机制,同时记录人为的数据拼接行为并利用机器学习等方法,不断增强平台智能的数据聚合能力。

(2)数据挖掘与模式发现。主要包括对数据的使用和内容挖掘两个方面,即:①数据使用 关注研究数据自身的价值和利用效率,类似于论文,建立计量评价模型(数据完整性、数据引证等),为研究人员寻找高价值的数据资源提供支持;②数据内容 关注来自不同领域数据、多个数据集的分析、比较及其知识发现,如基于相似数据集的隐性研究团队识别、基于数据质量(或一致性)的学术诚信识别等,平台应为上述活动的开展提供支撑。

3　人文社科研究数据管理平台关键技术实现

当前,已涌现出如 Dataverse、Dspace、EPrints、Fedora、Nesstar 等诸多平台原型,它们大多以数字资产系统为原型,侧重对数字资产的保存和管理,虽对数据分析和可视化功能支持较弱,但已为机构针对性的数据管理平台建设提供了较完整的解决方案,由此帮助我们摆脱了繁杂重复的基础通用功能开发而专注于对数据资源开发利用的功能开发。值得一提的是基于开源软件,二次开发构建数据管理平台,在国内外高校中也是比较普遍[33]。

作者以 Dspace 为平台原型,进行个性化设计与二次开发,继而完成人文社科研究数据管理平台的基础功能开发。其中,Dspace 软件最初是由美国麻省理工学院图书馆和美国惠普公司实验室合作开发并于 2002 年 10 月投入使用,它以内容管理发布为目标的数字资源存储系统,可实现对各种格式数字资源的收集、存储、索引和发布,具有完善的用户界面、可定制性强、较好的扩展性等,支持二次开发,目前已经被全球超过 300 家的机构在线使用,拥有众多用户和成功案例。文中开发的平台(由姚占雷负责平台的具体设计、谷俊负责平台的具体开发实现,访问地址 http://222.204.

246.126/rdmp/）是基于 Dspace 4.9 进行二次开发，开发环境为 springboot 2 开发框架、openjdk 1.8、postgresql 9.6、solr 7.3，且在开发过程中对基本的数据项、参数进行了重构，以满足前端页面内容自定义、数据需求管理（见图 8a）、数据关联研究成果（见图 8b）、数据申请授权管理（见图 9）等。

首页 / 个人中心 / 需求管理

数据空间管理
数据集合管理
数据条目管理
申请授权管理
通知列表管理
需求列表管理
站点设置
基本信息
修改密码

需求管理

姓名	研究方向	工作单位	办理状态	提交时间	操作
test	test	test	已拒绝	2020-01-07	查看详情 备注
彭思琪	公共政策	**科技情报研究所	待办理	2020-01-07	查看详情

‹ 1 ›

a. 数据管理平台后台功能示意

北美新闻记者招聘信息数据

所属数据空间/数据集合：科技人才/求职招聘

基础信息

标题	北美新闻记者招聘信息数据
作者	ThunderYao
学科	图书馆、情报与档案管理/情报学
简介	当前数据时间范围为2016.10-2019.12，数据记录数为23,392条，本数据条目不定期更新。
发布时间	2020-01-10 00:00:00
原始出处	http://www.journalismjobs.com
类型	数据集

展开

数据文件

文件名称	文件来源	描述	操作
北美新闻记者招聘信息数据	附件上传	暂无	数据文件下载 数据预览分析

相关成果

成果名称	出版刊物	发表时间	操作

b. 数据管理平台前台数据浏览示意

图 8　基于 Dspace 二次开发后的数据管理平台运行图（部分）



图 9　数据管理平台后台申请授权管理示意

3.1 开放互联的数据共享

数据共享包括了平台自身与其他数据平台的数据共享，以及各个平台上不同用户之间的数据文件共享。

平台自身与其他数据平台的数据共享，多是针对数据的题录信息交换，属于平台级应用，实现研究数据的最大化共享传播。平台采用 OAI-PMH 元数据收割协议进行两个独立平台之间的数据共享，前提是两个平台之间已经建立信任关系，同时平台支持 OAI-PMH 协议，允许其他数据平台进行元数据收割，如图 10 所示：



图 10　平台元数据收割参数配置

当数据被收割到其他数据平台上后，即可在该平台上实现用户间的数据共享。而为了保障数据提供者的合法权限，防止数据被用户滥用，平台在用户之间共享数据时，需要数据提供者对数据的使用进行授权，从而保障数据权属。目前，平台采用邮件和系统两种方式进行授权，当用户对数据有使用和下载需求时，系统会自动向提供者的邮箱发送一封请求授权的邮件，同时在数据提供者的系统管理后台同步，当且仅当数据提供者进行了数据授权，使用者才拥有数据使用的权限，见图 9。

同时，平台还对 Hyperledger Fabric 所支持的数据库进行了改进，用关系型数据库的存储方式替代传统超级账本的键值对数据存储方式（见图 11），以提升链上数据的查询处理能力，并据此在平台上预留了区块链基础设施的接口，当平台需要接入区块链基础设施时，可以在保障平台持续稳定运行的基础上，与区块链基础设施无缝对接。

```
flp :=blockIdxInfo.flp    //初始化，创建一个区块处理对象 flp。
txOffsets :=blockIdxInfo.txOffsets //初始化，创建一个交易索引信息对象 txOffsets。
txsfltr :=ledgerUtil.TxValidationFlags(blockIdxInfo.metadata.Metadata[common.BlockMetadataInde
x_TRANSACTIONS_FILTER] //对数据的有效性进行验证。
batch :=DB() //  初始化，创建关系型数据库对象 batch。
flp :=flp.json() //将 order 服务器传递过来的 JSON 格式数据转换为 flp 对象，用于数据库存储。
if_, ok :=index.indexItemsMap[blkstorage.IndexableAttrBlockHash]; ok {
        batch.Insert(constructBlockHashKey(blockIdxInfo.blockHash), flp)
}//调用 Insert 方法，将 flp 对象中的结果映射为数据库字段结构，并写入数据库。
```

图 11　链上数据结构化改造的核心代码

3.2 自助分析的数据探索

为进一步增强现有数据管理平台中的数据分析与可视化功能、支撑研究人员数据探索活动，平台借鉴商业报表思路，在统一的分析视窗（见图 12）下提供多维、动态交互的报表分析可视化功能，支持对可结构化的数据文件进行灵活自由的个性化数据探索。其中，为赋予报表分析的普适性，当前平台主要针对数据文件的数据项及取值进行通用挖掘分析，主要包括两类功能：①数据完整性、数据缺失值与极值、数据项分组、时间序列等库表结构类通用分析；②关键词云、实体抽取等科研活动中基础的文本分析。由此可见，此类功能在增强平台可视化分析能力的基础上，还面向研究人员提供简洁易用、一站式的数据分析服务，继而提升研究人员使用平台意愿，盘活与扩充数据资源。

通用报表分析是平台的一个智能分析模块，该模块可自动扫描并识别数据文件中相关字段及内容，并根据后台定义的数据分析模块，自动生成相应的可视化图表。其中，数据总量描述该数据文件中所包含的记录总量，如果发现该字段中的数据记录不完整，则提取该字段为缺失字段，此外还对所有字段数据类型的值进行对比，找出其中最大值和最小值字段，并标记为极值字段。同时在进行字段扫描过程，系统利用正则表达式进行字段类型的判定，如果该字段不符合数字字段或时间字段的匹配标准，则认定该字段为文本字段，并对于文本类型的字段，系统使用“结巴分词”工具进行文本的分词处理，并结合 TextRank 算法对字段内容进行标签提取，最终绘制出该数据文件的主题云图，便于用户深度挖掘数据集的主题。

同时，平台已与 Transwarp 大数据平台深度集成（集成的代码示意，见图 13），借助第三方平台丰富的分析工具，满足学者对数据资源开展更为专业的分析挖掘活动的需要，见图 14。

3.3 开发利用的数据增值

人文社科数据具有高度可复用性、持续产生价值的特点，同一研究方向的社科数据可以被多个研发团队复用，这使得多源数据资源拼接、聚合的价值凸显，平台建立起多源数据聚合的集成平台，重点面向平台内部结构化的数据资源，读取待聚合的数据资源特征项，为学者提供一套自由抽取与组合、内容灵活编辑的在线工具（见图 15）。同时，还支持接入学者的自有数据（excel、sql 等文件格式）。

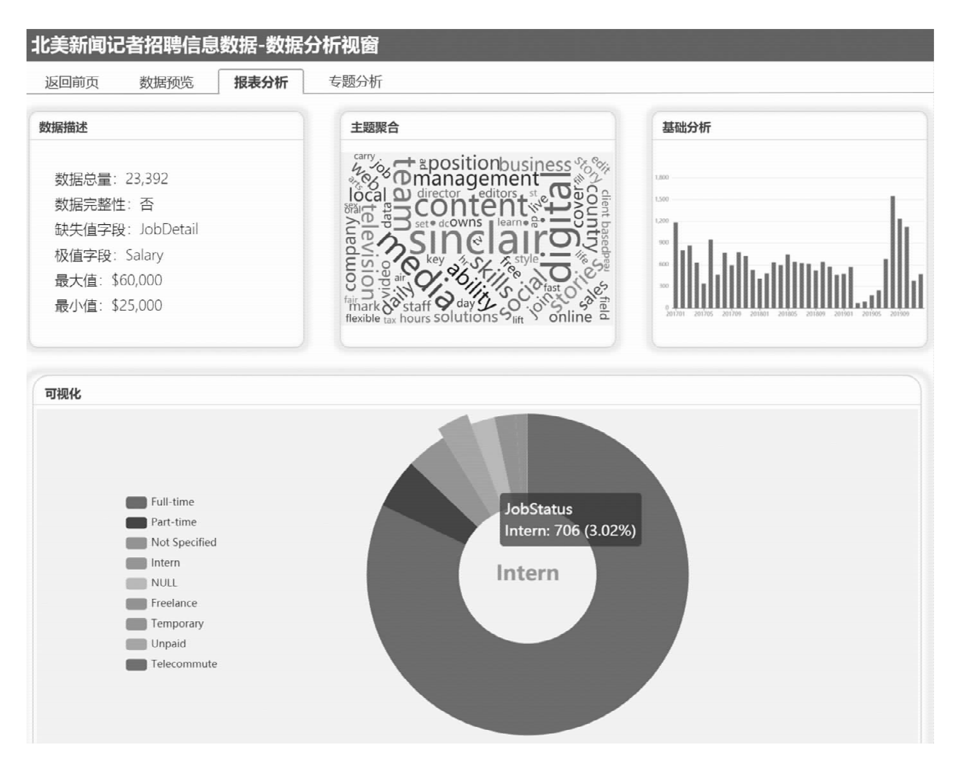


图 12　通用报表分析示意

```
private static String driverName = "io.transwarp.jdbc.InceptorDriver";
        public static void main(String[] args) throws SQLException {
                try {
                        Class.forName(driverName);
                } catch (ClassNotFoundException e) {
                        e.printStackTrace();
                        System.exit(1);
                }
                Connection conn =
DriverManager.getConnection("jdbc:hive://172.30.8.230:31326/;guardianToken=BvpBq6HQT6aVbDLjewreUcuz
W5HR.TDH");
                System.out.println("ok");
                Statement statement = conn.createStatement();
                 is_user = t.get_usr(username,userpwd);//账号验证
                 if (!is_user){
                        System.exit(1); //如果在数据系统中账号密码不匹配，则退出连接。
                }
                ResultSet recordSet = statement.executeQuery("show databases");
                ResultSetMetaData resultMeta = recordSet.getMetaData();
                int size = resultMeta.getColumnCount();
                while (rs.next()) {
                        StringBuffer value = new StringBuffer();
                        for (int i = 0; i < size; i++) {
                                value.append(rs.getString(i + 1)).append("\t");
                        }
                        System.out.println(value.toString());
                }
              recordSet.close();
              statement.close();
              conn.close();
```

图 13　接入 Transwarp 大数据平台的核心代码



图 14　接入第三方分析平台示意

多源数据聚合平台

⊕ 新增数据资源

☑资源名|马蜂窝游记
☐ 标题
☑ 出行天数
☑ 发表时间
☑ 人均费用
☑ 正文
☐ 作者

☑资源名|携程游记
☐ 标题
☐ 出行时间
☑ 发表时间
☑ 和谁
☐ 去哪
☑ 人均费用
☑ 天数
☐ 玩法
☑ 正文
☐ 作者

将选中的数据3条，生成新的数据集

☑ 导出	出行天数	发表时间	和谁	人均费用	正文	来源
☐		2020-02-07 18:16			童年中的	马蜂窝游记
☑		2020-02-07 11:20			一、因果定律世界上没有一件事是偶然发生的，	马蜂窝游记
☑	2 天	2020-2-7	一个人	1000 元	“滴滴、滴滴”，闹钟响了，伸手关闹钟后，我	携程游记
☐	1 天	2020-02-06 19:34			武康路位于	马蜂窝游记
☐	3 天	2020-02-05 16:55		1500RMB	2019.3.30我们去	马蜂窝游记
☐	7 天	2020-02-03 17:05		3000RMB	深秋已至，可	马蜂窝游记
☑	9 天	2020-02-02 23:03			♡写在开头的废话~	马蜂窝游记
☐	3 天	2020-01-03 22:20		1000RMB	写在前面	马蜂窝游记
☐		2019-12-29 18:20			2019/04/29-2019/05/03当	马蜂窝游记
☐	6 天	2019-12-22	亲子	8000 元	新加坡萌版的道路一直在和我们捉迷藏。明明在	携程游记
☐	1 天	2019-10-19	亲子		★旅游地点：张家港永联小镇永联小镇位于永钢	携程游记
☐	3 天	2019-9-25	一个人		写在前面的话	携程游记
☐	5 天	2019-09-17 11:19		2900RMB	成都到南京到上海，河海大学至外滩至迪士尼	马蜂窝游记
☐	1 天	2019-6-17	和朋友		余秋雨是上海的，上海的古迹没有多少好看的，	携程游记
☐	6 天	2019-6-2	和朋友	3000 元	向北。今年的夏天里，我仿佛做了一个很长的梦。	携程游记
☐	3 天	2019-5-30	一个人	1500 元	上海，是我非常喜欢的城市！从我的第一份工作	携程游记
☐	2 天	2019-5-21	夫妻	500 元	心驰“同里”心同理“琴笙楠舍”情难舍——“	携程游记
☐	7 天	2019-5-8	和朋友	6000 元	我们真的要过了很久很久，才能够明白，自己会	携程游记
☐	5 天	2019-05-05 15:38		9000RMB	..	马蜂窝游记
☐	5 天	2019-4-24	一个人	5000 元	生命里总有那么一个人惊艳了时光令人念念不忘	携程游记

上一页　当前页 1/2,000　下一页

图 15　多源数据聚合平台示意

4　结语

有别于当前主流的研究数据管理平台架构侧重对数据的科学管理，本文主要拓展与丰富了人文社科研究数据管理平台的功能与定位，能够进一步充实或完善相关研究与工作实践，旨在促进人文社科研究研究数据的共享与重用活动。

具体而言，本文主要围绕数据资源的多途径采集与规范、自助分析与开发利用，提出了面向全生命周期的人文社科研究数据管理平台基础框架，据此对数据管理平台的功能有针对性地进行了优化重组，注重对数据资源的二次开发利用，如：自助报表的分析视窗，满足人文社科学者科研活动中常见的文本计算与统计分析需要；多源数据拼接功能，满足人文社科学者对同一主题不同数据集的数据资源聚合，等等。

而针对高价值的数据资源，在取得数据所有者授权的前提下，平台亦可以开展系列专项增值活动，如按照主题、学科、事件等形式加工汇编成特色专题数据进行出版发行，并充分利用数据在平台上的使用与评价情况，为数据出版等工作提供更为丰富的支撑；结合数据资源自身的研究属性、平台富含的在线工具等，可为学科领域内科研新人搭建交流平台，助力其重现经典

研究活动、了解研究范式，以快速把握相关研究路径。这些新型尝试，将为进一步推动研究数据资源的开发利用提供新范式、新路径，并成为平台后续优化升级所关注的重点。

参考文献：

[1] 孙玉伟，成颖，谢娟. 科研人员数据复用行为研究：系统综述与元综合[J]. 中国图书馆学报，2019(5)：110－130.

[2] 王晓光. 加强人文社科数据资源建设与管理[N]. 光明日报，2018－07－05(11).

[3] 张亚楠，黄晶丽，王刚. 考虑全局和局部信息的科研人员科研行为立体精准画像构建方法[J]. 情报学报，2019，38(10)：1012－1021.

[4] 钱锦林，刘贵峰. 国外科研数据管理研究综述[J]. 情报理论与实践，2017，40(10)：130－134.

[5] 邢文明，杨玲. 我国科研机构科研数据管理现状调研[J]. 数字图书馆论坛，2018(12)：27－33.

[6] 胡永生，刘颖. 基于用户调查的高校科学数据管理需求分析[J]. 图书情报工作，2013，57(6)：28－32＋78.

[7] 王丹丹. 科学数据管理服务需求识别方法研究[J]. 大学图书馆学报，2018，36(1)：41－47.

[8] BALL A. Review of Data Management Lifecycle Models[EB/OL]. [2019－09－26]. http://opus. bath. ac. uk/28587/.

[9] 钱鹏. 信息生命周期管理两重性辨析：以科学数据管理为例[J]. 情报理论与实践，2013，36(3)：11－14.

[10] DARLINGTON M，BALL A. A Research Data Management Plan for Engineering Research[EB/OL]. [2019－09－30]. http://opus. bath. ac. uk/30104/.

[11] RICE R，HAYWOOD J. Research data management initiatives at University of Edinburgh[J]. International journal of digital curation，2011，6(2)：232－244.

[12] 陈大庆. 国外高校数据管理服务实施框架体系研究[J]. 大学图书馆学报，2013(6)：10－17.

[13] 张培凤，张连分. 全球科研范式变革下的图书馆科学数据管理服务创新——基于数据管理生命周期的视角[J]. 图书馆理论与实践，2019(5)：39－48.

[14] 周玉琴，邢文明. 我国科研数据管理与共享政策体系研究[J]. 中华医学图书情报杂志，2018，27(8)：1－7.

[15] 何青芳. 国外科学数据管理政策的调查与分析[J]. 上海高校图书情报工作研究，2016，26(2)：9－13.

[16] 姜鑫. 科学数据开放政策研究现状分析及未来研究动向评判[J]. 现代情报，2016(2)：167－171.

[17] 邢文明，汤雅静，秦顺. 国外教育机构科研数据管理政策大纲解读及启示[J]. 数字图书馆论坛，2019(5)：9－16.

[18] 鄂丽君. 国外大学图书馆的科研数据管理教育[J]. 情报资料工作，2014(1)：101－105.

[19] 张艳梅. 用户数据素养教育视角下的图书馆科学数据管理研究[J]. 图书与情报，2015(4)：139－141，109.

[20] 崔宇红，李伟绵. 研究数据管理进展评述[J]. 图书馆杂志，2017(1)：12－19.

[21] 柴会明，张立彬，赵雅洁. 国内图书馆科学数据研究述评[J]. 图书情报工作，2019，63(7)：116－126.

[22] 丁宁，马浩琴. 国外高校科学数据生命周期管理模型比较研究与借鉴[J]. 图书情报工作，2013，57(6)：18－22.

[23] 井润田，王蕊，周家贵. 科研团队生命周期阶段特点研究——多案例比较研究[J]. 科学学与科学技术管理，2011(4)：173－179.

[24] 贾玉文，李超群. 嵌入科研生命周期的数据资源整合模型研究[J]. 图书馆学刊，2019(2)：51－55.

[25] 李伟绵. 基于生命周期理论的研究数据管理服务评估研究[D]. 北京：北京理工大学，2016.

[26] CROWSTON K，QIN J. A capability maturity model for scientific data management：evidence from the literature[J]. Proceedings of the American Society for Information Science and Technology，2011，48(1)：1－9.

[27] MARCHIONINI G，杨冠灿，芦昆. 科研数据管理：保障数据质量，促进 iSchools 新科学研究[J]. 图书情报知识，2013(4)：4－9.

[28] 崔海媛. 研究数据管理和服务指南[M]. 北京：海洋出版社，2019.

[29] Data Management General Guidance[EB/OL]. [2020－06－25]. https://dmptool. org/general_guidance/.

[30] SAYRE F D，BAKKER C J，JOHNSTON L R，et al. Where in academia are ELNs? Support for electronic lab notebooks at top American research universities[C]//Poster presented at the Association of College & Research Libraries Conference. Baltimore：ACRL，2017.

[31] 崔旭，赵希梅，王铮，等. 我国科学数据管理平台建设成就、缺失、对策及趋势分析——基于国内外比较视角[J]. 图书情报工作，2019，63(9)：21－30.

[32] 谷俊，许鑫. 人文社科数据共享模型的设计与实现——以联盟链技术为例[J]. 情报学报，2019(4)：354－367.

[33] 洪正国，项英. 基于 Dspace 构建高校科学数据管理平台——以蝎物种与毒素数据库为例[J]. 图书情报工作，2013，57(6)：39－42，84.

作者贡献说明：

姚占雷：平台设计与文稿撰写；

谷俊：平台开发与文稿修改；

许鑫：总体设计与文稿定稿。

Design and Implementation of Management Platform for Humanities and Social Sciences Research Data in the Perspective of Full Life Cycle

Yao Zhanlei[1]　Gu jun[2]　Xu Xin[1,3]
[1] Faculty of Economics and Management, East China Normal University, Shanghai 200062
[2] School of Humanities, Shanghai Normal University, Shanghai 200233
[3] Social Survey and Data Center, East China Normal University, Shanghai 200241

Abstract: [**Purpose/significance**] Policies that giving clear guidance and regulations on the management, sharing and utilization of scientific data have been issued by China in recent years. However, architectures of current data management platforms focus on the scientificity of data management, not the efficiency of data sharing and utilization. Based on systematically expanding the data management functions of existing platforms, this study focused on solving the problems of sharing and using scientific data. [**Method/process**] Based on the investigation and literature review, this study first clarified the necessity and predicament of the construction of a management platform for humanities and social sciences research data. Second, from the perspective of full life cycle, the study systematically designed the core functions and explained the characteristics of the platform. Then, the study described the details of the realization of key functions by a real case. [**Result/conclusion**] Aimed at open and interconnection, development and utilization, and self-service analytics, a basic framework of research data management for full life cycle was established. Based on the framework, a management platform for humanities and social sciences research data was actualized. The platform can be a characteristic case and a reference for related practices and studies.
Keywords: data management　humanities and social sciences　development and utilization　platform construction